

Research Article

Markers or metagenomes: sequencing marine eukaryotic DNA for better biodiversity surveys

Nastassia V. Patin^{1,2}, Kathleen J. Pitz³, Jacoby D. Baker³, Francisco P. Chavez³, Kelly D. Goodwin⁴¹ Integrated Oceanography Division, Scripps Institution of Oceanography, UC San Diego, La Jolla, CA, USA² Southern California Coastal Water Research Project, Costa Mesa, CA, USA³ Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA⁴ NOAA Ocean Exploration, La Jolla, CA, USACorresponding author: Nastassia V. Patin (nvpatin@ucsd.edu)

Abstract

Marker gene sequencing (“metabarcoding”) is the primary sequencing approach currently used for molecular biodiversity surveys, but this approach is taxonomically limited and hampered by amplification biases. In contrast, shotgun metagenomes avoid extensive PCR amplification and can theoretically capture the full taxonomic breadth of the eDNA pool. However, eukaryotic DNA is often a small component of metagenomes; consequently they have seen limited use in metazoan biodiversity surveys. Here, we compare metabarcoding and shotgun metagenomes on a large (>200 sample size) set of marine water column eDNA samples and show that metagenomes can provide biodiversity information comparable to that of metabarcoding surveys. They provide more consistent and even detections of important phytoplankton like chlorophytes and haptophytes and can detect metazoan taxa of management and conservation interest, such as the humpback whale. However, genus- and species-level taxonomic assignments are often inaccurate for metagenomic data and require more manual annotation relative to metabarcoding taxonomy. Reference database gaps remain an obstacle to accurate and comprehensive biodiversity surveys for both metabarcoding and shotgun metagenomes. We provide examples of taxa that may benefit from one approach over another and highlight cases of metagenomic utility.



Academic editor: Anastasija Zaiko

Received: 11 January 2025

Accepted: 8 June 2025

Published: 16 July 2025

Key words: Amplicon sequencing, bioinformatics, eDNA, metagenomics, reference databases

Citation: Patin NV, Pitz KJ, Baker JD, Chavez FP, Goodwin KD (2025) Markers or metagenomes: sequencing marine eukaryotic DNA for better biodiversity surveys. *Metabarcoding and Metagenomics* 9: e144935. <https://doi.org/10.3897/mbmg.9.144935>

Copyright: This is an open access article distributed under the terms of the [CC0 Public Domain Dedication](https://creativecommons.org/licenses/by/4.0/).

Introduction

The use of environmental DNA (“eDNA”) is increasingly recognized as a powerful approach for biodiversity assessments. Molecular methods bypass many of the limitations associated with visual, auditory, or other survey methods, including sampling time and effort and the detection of organisms not immediately present at the time of sampling. In the past two decades, the application of eDNA sequencing and analysis has expanded from microbial communities to include higher trophic level organisms including phytoplankton, zooplankton, and even vertebrates (Thomsen and Willerslev 2015; Deiner et al. 2017; Ruppert

et al. 2019; Compson et al. 2020; Takahashi et al. 2023). The shift of attention from mostly single-celled organisms like bacteria and plankton to multicellular metazoans has introduced new challenges and amplified existing ones, such as marker gene and primer selectivity (e.g. Cowart et al. 2015; Zhang et al. 2020; Casey et al. 2021) and estimating the relationship of sequence read numbers to target organism abundance (Harrison et al. 2019; Pont et al. 2023; Shelton et al. 2023). A microbes-to-mammals marker gene approach would provide a scalable, cost-effective biomonitoring method, but is hindered by the lack of truly “universal” primers. The potential utility of large-scale eDNA monitoring for management and conservation has spurred interest in addressing these challenges and finding solutions.

Marker gene sequencing, also known as amplicon sequencing or metabarcoding, is by far the most common approach for eukaryotic biodiversity assessments (Takahashi et al. 2023). In this approach, clade- or taxon-specific primers target organisms of interest and all amplified marker genes are sequenced to provide an overview of target taxa or even, in some cases, genetic haplotypes within species (Stat et al. 2017; Tsuji et al. 2020; Andres et al. 2023). However, the taxonomic breadth of the survey depends on the specificity of the primers, which may exclude certain taxa within the target group (Tremblay et al. 2015; Clarke et al. 2017; Alberdi et al. 2018; Bylemans et al. 2018; Serite et al. 2023). Moreover, the marker gene in question may not be sufficient to distinguish among all targeted species, for instance if two species have the exact same amplicon sequence. Conversely, a single nucleotide difference in a primer can alter the list of detected taxa (Apprill et al. 2015; Piñol et al. 2015). Such primer or other amplification biases during the marker gene amplification step (Wintzingerode et al. 1997) can prevent the detection of rare taxa or distort read levels relative to their representation in the eDNA pool (Elbrecht and Leese 2015; Piñol et al. 2015; Kelly et al. 2019; Serite et al. 2023). These biases can be species-specific and are therefore challenging to calculate for an entire group of target taxa.

Metagenome shotgun sequencing is an approach that generates sequence data for the entire eDNA pool, including but not limited to marker genes. Shotgun library preparation includes minimal PCR amplification, if any, and thus avoids the amplification biases associated with metabarcoding (Tringe and Rubin 2005). While metagenomes may contain orders of magnitude more sequence data than metabarcoding data sets, they are typically dominated by prokaryotic DNA and may contain only low levels, if any, of target eukaryotic organism sequences (Tedersoo et al. 2015; Stat et al. 2017). However, some studies have successfully captured ecologically relevant eukaryotic sequences with shotgun metagenomics (Cowart et al. 2018; Singer et al. 2020; Machida et al. 2021; McCauley et al. 2024), and such data can be highly effective for eukaryotic protist genomics (Delmont et al. 2022). Comparisons of the two approaches for animal diet studies, where prey DNA makes up a large fraction of the DNA pool, have shown metagenomes provide higher taxonomic diversity and resolution of target organisms but can miss rare taxa (Srivathsan et al. 2015; Serite et al. 2023).

Regardless of which sequencing method is used to assess eukaryotic biodiversity, the taxonomic results depend on the completeness and accuracy of the sequence reference database used to assign taxonomy. Several

comprehensive and manually curated reference databases exist for eukaryotic marker gene sequences. These databases can include sequences for one gene in one taxonomic group (e.g. PR2, 18S rRNA / protists; Guillou et al. 2012) or for multiple genes (e.g. MetaZooGene, cytochrome oxidase 1 (CO1) and 18S / 12S / 16S / 28S rRNA; Bucklin et al. 2021). However, unlike the 16S rRNA marker gene database SILVA that encompasses all prokaryotes, there is no manually curated, comprehensive eukaryotic marker gene database that includes protists, invertebrates, and higher vertebrate taxa. NCBI-nt (nucleotide), of which GenBank records the primary sequence content, remains the most comprehensive repository of eukaryotic sequence information (Strasser 2008; Arita et al. 2021; Sayers et al. 2023; Renner et al. 2024) and taxonomy of mitochondrial genes is highly accurate (Leray et al. 2019). However, NCBI has relatively little manual curation and is subject to misannotation (Keck et al. 2023). Moreover, it uses a taxonomic format that is incompatible with many other biodiversity taxonomy codes (e.g. the World Registry of Marine Species). These issues also apply to shotgun metagenome taxonomic annotation, which is generally performed using NCBI-nt or a derivative thereof (e.g. UniRef, UniProt). If NCBI or a related database is used for both marker gene and metagenome annotation, differences in gene representation will influence the results. While annotation of metabarcoding data relies on the presence of marker gene sequences, metagenomic functional gene annotations rely largely on whole genome sequences. Some taxa may have been extensively barcoded, while others (e.g. organisms of interest in biomedical or agricultural research) may have more whole genome sequence data and thus the annotations of metagenomes will be biased towards these taxa. The impact of database composition on sequence annotation in the context of marine biodiversity is an important knowledge gap with implications for sequencing choice and data interpretation.

Given the ever-decreasing cost of high-throughput sequencing and the many factors affecting eukaryotic sequence annotation, the time is ripe for an assessment of optimal sequencing approaches for eukaryotic eDNA. Here, we compare shotgun sequencing with metabarcoding for capturing eukaryotic biodiversity in marine water column samples. Two different marker genes were used to assess metabarcoding efficacy: 18S ribosomal RNA (“rRNA”) and cytochrome oxidase 1 (“COI”). These markers encompass a wide range of eukaryotic diversity which, while not appropriate for targeted surveys of metazoans such as teleost fish due to high levels of off-target amplification (Collins et al. 2019), provide a better comparison with metagenomes for a microbes-to-mammals survey goal. Other studies have addressed the differences between these two genes (e.g. Clarke et al. 2017; Casey et al. 2021); here, we focus on the differences between shotgun metagenomic sequencing and each of the markers. We do not discuss the influences of sample collection or processing despite small differences among the sample sets. On a per-sample basis, only the sequencing approach varied and thus we could directly compare the results of metagenomic and amplicon sequencing in aggregate. We address 1) taxonomic composition (quantitative; relative abundances), 2) taxonomic diversity (qualitative; presence/absence of different taxa), and 3) the effect of reference database composition on taxonomic assignment discrepancies.

Methods

Sample collection

eDNA sampling was performed in spring and fall of 2018 by the Monterey Bay Aquarium Research Institute as part of their Controlled, Agile, and Novel Observing Network (CANON) Initiative. Samples were also collected on the NOAA Rockfish Recruitment and Ecosystem Assessment Survey (RREAS) in the spring of 2018 and 2019. CANON eDNA samples were collected as described in Truelove et al. 2022, and RREAS samples were collected as described in Patin and Goodwin 2022. Briefly, marine water column samples from the California Current Ecosystem (CCE) were filtered onto 0.22 μm filters using either an autonomous underwater vehicle equipped with an Environmental Sample Processor (Yamahara et al. 2019), or shipboard using Niskin bottles and a peristaltic pump. Sample collection information and environmental metadata are provided in Suppl. material 1: table S1.

Sample processing, library preparation, and sequencing

DNA was extracted from the CANON 2018 filters at MBARI according to methods described in Truelove et al. 2022, from the RREAS 2018 filters at NOAA's Southwest Fisheries Science Center according to methods described in Closek et al. 2019, and from the RREAS 2019 filters at Oregon State University's Center for Quantitative Life Sciences Core Facilities ("OSU") according to Patin and Goodwin 2022. Briefly, CANON and RREAS 2018 DNA was extracted using the Qiagen DNEasy Blood & Tissue Kit with some modifications including the addition of a bead-beating step to maximize yield (CANON 2018: dx.doi.org/10.17504/protocols.io.xjufknw, RREAS 2018: dx.doi.org/10.17504/protocols.io.yrqfv5w). RREAS 2019 DNA was extracted with the Omega Bio-Tek Mag-Bind blood and tissue DNA HDQ 96 kit with some modifications and a KingFisher instrument (dx.doi.org/10.17504/protocols.io.14egn6rxql5d/v1).

The 18S rRNA and COI marker gene libraries were amplified using the 18S V9 (Amaral-Zettler et al. 2009; Stoeck et al. 2010) and COI forward (Leray et al. 2013) and reverse (Folmer et al. 1994) primer set, respectively. CANON and RREAS 2018 amplicon samples were amplified at MBARI (18S V9 amplification protocol: dx.doi.org/10.17504/protocols.io.36wgq3d8olk5/v2, COI amplification protocol: DOI dx.doi.org/10.17504/protocols.io.4r3l2q1ejl1y/v1) and sequenced at the Michigan State University Research Technology Support Facility Genomics Core ("MSU"). Library preparation was performed using Fluidigm adapters, normalized with the Applied Biosystems™ SequelPrep™ Normalization Plate Kit, and sequenced on an Illumina MiSeq using 2 × 150 bp paired-end chemistry for 18S V9 amplicons and 2 × 250 bp chemistry for COI amplicons. RREAS 2019 amplicons were amplified and sequenced at OSU using MSU protocols described above with the following steps modified: 1) Primary PCR amplification – Illumina Nextera sequencing adaptors replacing the Fluidigm sequencing primers; 2) Secondary PCR amplification – Illumina NexteraXT v.2 primer set for indexing samples. Thermal cycling program for indexing: Initial denaturation 95 °C for 3 minutes; 8 cycles of: denaturation at 95 °C for 30 seconds, annealing at 55 °C for 30 seconds, extension at 72 °C for 30 seconds; final extension at 72 °C for 5 minutes; Hold at 4 °C; 3) Normalization – amplicon

products were quantified with a fluorescent plate reader and Quant-iT 1x dsDNA HS Assay then we normalized manually using an EpMotion liquid handler. All amplicon raw FASTQ files are available on the NCBI Sequence Read Archive (BioProject number PRJNA726280).

The CANON 2018 metagenomes were prepared and sequenced by Laragen Inc. (Culver City, CA). The RREAS 2018 metagenomes were prepared and sequenced at MSU. Both 2018 sample sets were sequenced using an Illumina NovaSeq with 2 × 150 bp chemistry. The RREAS 2019 metagenomes were prepared and sequenced by CosmosID (Germantown, MD) as described in Patin and Goodwin 2022 using an Illumina HiSeq with 2 × 150 bp chemistry. Metagenome and amplicon sequencing statistics are provided in Suppl. material 2: table S2. All raw metagenome reads are available on NCBI (BioProject number PRJNA853328).

Initial sequence data processing

Data samples (218) were selected such that every eDNA sample had data for all three sequencing approaches.

Raw metabarcoding reads were processed through a custom shell script adapted from the Banzai pipeline and run on a per-plate basis (<https://github.com/MBARI-BOG/BOG-Banzai-Dada2-Pipeline>). Within the script, primer sequences were first removed from FASTQ files through the program Atropos (Didion et al. 2017). FASTQ files were then processed with DADA2 (Callahan et al. 2016). Within DADA2, reads were trimmed to remove low-quality regions and filtered by quality score, sequencing errors were modeled and removed, and reads were then merged and chimeric sequences removed.

Quality control of raw shotgun metagenome reads was performed with multitrim (<https://github.com/KGerhardt/multitrim>), a wrapper for FaQCs (Lo and Chain 2014), fastp (Chen 2023), and Falco (De Sena Brandine and Smith 2021) to detect and remove Illumina adapters and trim reads based on quality scores using default parameters. Interleaved reads were assembled into contigs using metaSPAdes (Nurk et al. 2017) with default parameters.

Following read-based taxonomy assignment (see below), RREAS2019 metagenomes were found to contain human DNA contamination (Patin and Goodwin 2022). These reads were decontaminated using BMTagger (<ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/>) which removed reads mapped against a human reference genome (GRCh38_latest_genomic). RREAS2019 metagenomes were re-annotated and assembled following this process.

Taxonomy assignment

Metabarcoding taxonomy was assigned to the ASVs through blastn searches to NCBI GenBank's non-redundant nucleotide database (nt) (Camacho et al. 2009; NCBI Resource Coordinators et al. 2018). The following BLAST parameters were used: percent identity = 80, word size = 11/20 (COI/18S), e-value = 1e-5, maximum target matches = 100. BLAST results were filtered using MEGAN6's lowest common ancestor (LCA) algorithm (Huson et al. 2016). Only hits with an e value ≤ 1e-25, ≥100 bitscore, and whose bitscores were within the top 2% of the highest bitscore value for each ASV were considered by MEGAN6. The MEGAN6 parameter LCA percent was set to 80%.

Post-MEGAN6 filtering was performed to ensure only ASVs with a hit of $\geq 97\%$ sequence identity and ≥ 200 bitscore were annotated to the species level. Only ASVs with a hit of $\geq 95\%$ sequence identity and ≥ 150 bitscore were annotated to the genus level. Annotations were elevated to the next highest taxonomic level for ASVs that failed those conditions.

Metagenomes were taxonomically annotated in two different ways. First, un-assembled reads were assigned with sourmash (Pierce et al. 2019) and the associated branchwater plugin using a database composed of GTDB prokaryotic genomes (Parks et al. 2022) and GenBank eukaryotic genomes (“RocksDB”; <https://github.com/sourmash-bio/sourmash/issues/3504>). Sourmash sketches (compressed representations of sequence data) were generated with a k-mer size of 51 using ‘sourmash scripts manysketch’ with the ‘-p abund’ parameter to weight k-mers by abundance. Signatures were classified with ‘sourmash scripts fastmultigather’ and annotated with ‘sourmash tax metagenome.’ The eukaryotic fractions were calculated as the proportion of base pairs belonging to the “superkingdom Eukaryota” The proportion of base pairs belonging to “Bacteria” and “Archaea” (“superkingdom” ranks) combined were classified as “prokaryotic” and “unclassified” superkingdom proportions were classified as such. The total number of base pairs assigned to each “superkingdom” was calculated using sourmash’s “f_weighted_at_rank” proportion multiplied by the total number of base pairs in the metagenome, which was calculated using the ‘stats’ function of the ‘seqkit’ toolkit (Shen et al. 2016). These calculations were performed in the Jupyter notebook `format_tax_compositions.ipynb`.

Second, assembled contigs were run against the UniRef 100 database (Release: 2024_01, version 24-Jan-2024) using mmseqs (Mirdita et al. 2021) with the “easy-taxonomy” command and the following parameters: `--lca-mode 3 --orf-filter 1 --orf-filter-s 3.5 -s 4 --tax-lineage 1`. The mmseqs easy-taxonomy annotations were filtered to exclude viruses and contigs with ‘no rank’ as assigned Taxid. Processing was performed using the custom Python script `filter_mmseqs_lca_taxonomy.py`. Annotations were quantified by mapping raw metagenome reads onto the contigs using coverM (Aroney et al. 2025) and minimap2 (Li 2018) with the coverage method (parameter “-m”) set to “tpm” to calculate mapping levels as transcripts per million (Li et al. 2010). Contig taxonomy and TPM values were linked and formatted for plotting in the custom Python script “`format_tax_compositions.ipynb`.”

Rarefaction curves

To assess whether sample diversity was adequately captured with the applied sequencing depth, rarefaction (accumulation) curves were generated for metagenome and amplicon data sets. K-mers were used as a proxy for taxonomic diversity in the metagenomes (Wood and Salzberg 2014) while individual ASVs were used for amplicons.

For metagenomes, the sourmash abundance-weighted k-mer signatures (“sketches”) were downsampled with a scaled value of 100,000 using ‘sourmash sig downsample’ to decrease the signatures sizes and associated processing time. The downsampled signatures were used to generate rarefaction curves with the R package `sourmashconsumr` (Chou and Reiter 2024). Code for rarefaction calculation and plotting was performed using the R script ‘`sourmash_rarefac.R`.’

Rarefaction curves for amplicons were generated using the `ggrare` function in the `Ranacapa` R library. The number of sequences in each sample was plotted in accumulation curves against the corresponding species richness in the script “Rarefaction_curves.R.”

Taxonomic composition and diversity comparisons

Percent composition of each taxonomic rank in each sample and across samples was calculated for all four data sets (unassembled metagenome reads, assembled metagenomic contigs, 18S amplicons, and COI amplicons) using the Jupyter Notebook `format_tax_compositions.ipynb`. Results were formatted for Krona (Ondov et al. 2011) using `format_krona_tsvs.ipynb` and Krona plots were generated using `ktImportText` in Krona. Because NCBI does not provide phylum-level annotation for organisms in the Class Dinophyceae (ie dinoflagellates), taxa with the class-level annotation were assigned the name “Dinophyceae” at the phylum to replace NCBI’s “unknown” designation. These proportions were grouped at the phylum and class levels, averaged across samples, and eukaryotic phyla were compared in the Jupyter notebook `taxa_comparison_stackedbars.ipynb`. Levels of the phylum Streptophyta in metagenomes were compared among depth groupings (0–7 m, 10–100 m, and integrated) in the Jupyter notebook `check_surface_samples.ipynb`.

Sample composition was directly compared between metagenomic contigs and amplicon data for each sample (paired sample comparison). The relationships for each phylum with a relative abundance above 0.1% in either data set were modeled using the `lm()` function in base R, and Pearson correlation coefficients (r) were calculated for each regression. These correlation coefficients were plotted for both 18S and COI marker genes. Class-level regressions were also plotted using the same function. Code for the regressions and correlation calculations can be found in `euks_asvs_mgns_regression.Rmd`.

Phylogeny of outlier Dinophyceae ASV

To investigate the identity of the single COI ASV present in high relative abundance (ASV_1416), all ASVs annotated as Dinophyceae across all samples were aligned using Muscle in AliView v. 1.28 (Larsson 2014). A maximum likelihood phylogeny was built using the Tamura-Nei model with uniform rates in MEGA (Tamura et al. 2021) and displayed in TreeViewer (Bianchini and Sánchez-Baracaldo 2024) as a rooted tree.

Phylogeny of Magnoliopsida 18S ASVs

All 35 18S ASVs annotated as belonging to the Family Magnoliopsida were aligned with five reference sequences downloaded from GenBank. The reference sequences included three Magnoliopsida species, *Micromonas pusilla* (Phylum Viridiplantae, same as Magnoliopsida), and *Acrodontium crateriforme* (Phylum Ascomycota) (Suppl. material 4: table S4). Reference sequences were first trimmed to the 18S-v9 rRNA gene region using a Muscle alignment generated with the metabarcoding primer sequences in AliView v. 1.28 (Larsson 2014). Trimmed sequences were then aligned with the ASVs using the same

approach and a maximum likelihood phylogeny was built using the Tamura-Nei model with uniform rates in MEGA (Tamura et al. 2021). It was displayed in TreeViewer (Bianchini and Sánchez-Baracaldo 2024) as a rooted tree.

Reference database comparison: marker gene vs whole genome sequences in NCBI

We compared the nucleotide content of marker genes with whole genome sequences for each phylum in NCBI to assess how availability of reference sequences may have affected annotations. To calculate genome sequence content for each phylum observed in the sequence data, downstream species taxon IDs were extracted using the R library taxizedb (Chamberlain et al. 2023). The taxon IDs were used to download genome AssemblyStats files from GenBank using the R library biomartr (Drost and Paszkowski 2017) in the R script getNCBIGenomeStats.R. Only one representative genome file per species was downloaded. Average genome sizes and total genome nucleotides were calculated using the AssemblyStats files in the Jupyter notebook parse_genome_assembly_stats.ipynb.

To calculate marker gene content, existing rCRUX databases were leveraged (COI: <https://zenodo.org/records/8407632>, 18S v9: <https://zenodo.org/records/8407878>). rCRUX uses primer sequences to extract marker gene sequences from an existing database using an initial seed-building step followed by a comprehensive seed BLAST job (Curd et al. 2024). Both rCRUX databases in question were built from NCBI nt. The taxonomy output of the rCRUX blast_seeds_output job was dereplicated using the bash command “sort | uniq” and the number of marker genes were manually counted using the bash command “grep | wc -l” for each phylum.

Results

Eukaryotic vs prokaryotic sequence composition

The two sequencing approaches generated datasets that differed substantially in quality and quantity of eukaryotic DNA sequences. Approximately 88% and 66% of the 18S and COI annotated metabarcoding data, respectively, were eukaryotic compared to 6.6% of shotgun data in assembled contigs and only 2% in the un-assembled reads. However, since metagenomes were 3–4 orders of magnitude larger than metabarcoding datasets, the average total number of eukaryotic nucleotides was still greater than the 18S data, while COI data had approximately twice the amount of eukaryotic sequence data (Table 1). In metagenomes, the fraction of unclassified sequences decreased after reads were assembled into contigs, with corresponding increases in all classified domains including Eukaryota (Fig. 1).

Table 1. Overview of the average sequence data set sizes and average eukaryotic composition, in base pairs, of the three data sets. The 18S amplicon and metagenome sequencing was performed with 2 × 150 bp paired-end chemistry while COI amplicons were sequenced with 2 × 250 bp PE.

Data set	Avg. # bps/sample	Avg. # eukaryotic bps/sample	Avg. % eukaryotic bps/sample
18S rRNA amplicon	11,940,775	10,577,628	88.6
COI amplicon	29,997,108	19,816,450	66.1
Metagenome reads	30,921,932,487	310,734,500	1.0

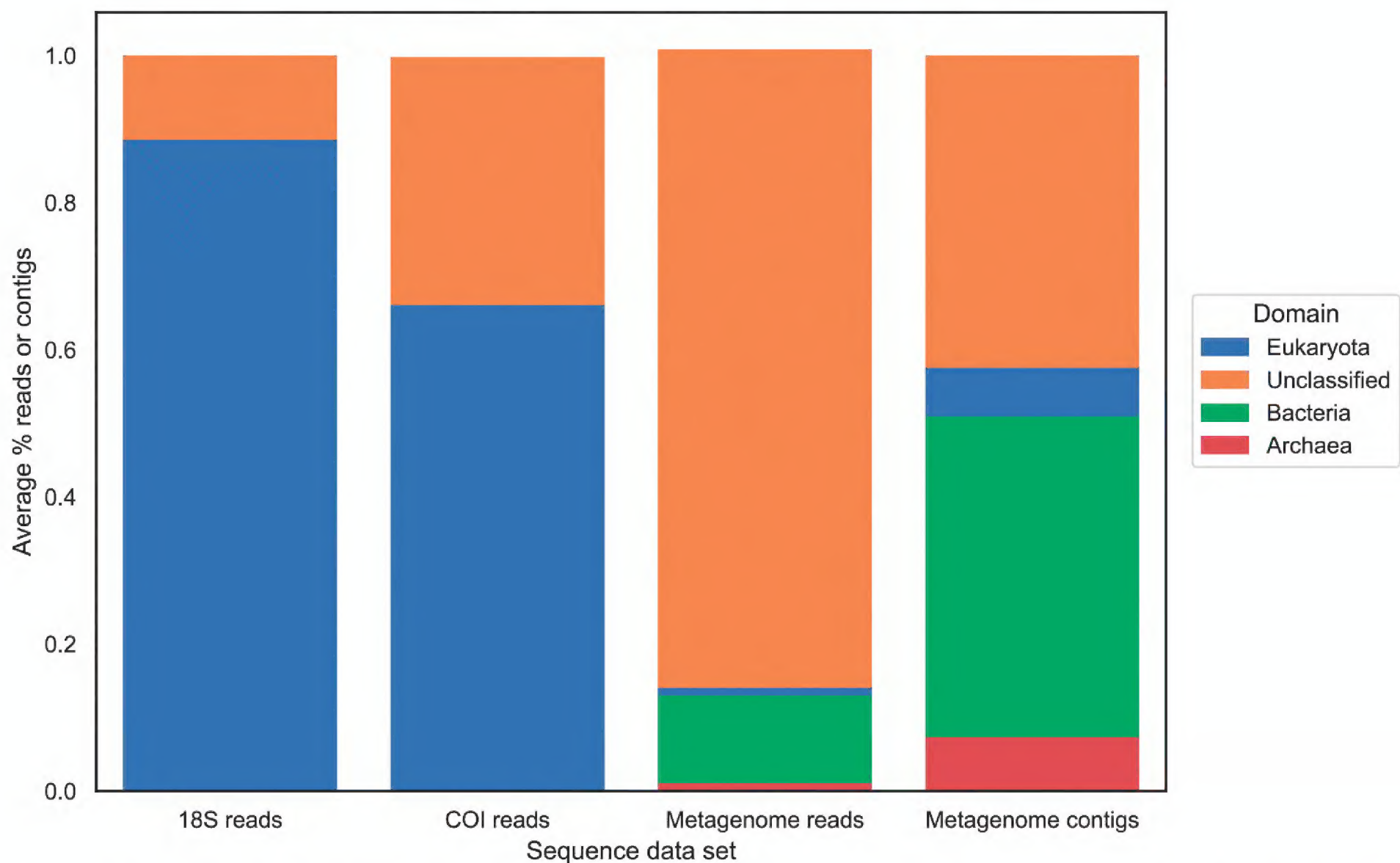


Figure 1. Average prokaryotic, eukaryotic, and unassigned taxonomic composition of marker gene and shotgun metagenome sequence data sets. Annotations of unassembled metagenomic reads are shown separately from annotations of assembled contigs. Values are presented as a % total base pairs, except for metagenome contigs which are shown as % contigs.

Rarefaction curves of each sequence data set showed most sample diversity beginning to plateau (Suppl. material 6: fig. S1), with most of the amplicon data showing minimal increase in diversity past ~40,000 sequences. The Lasker 2019 COI data set was a notable exception, with most samples likely requiring well over 100,000 sequences to approach saturation as demonstrated by the single sample with exceptionally high sequencing depth (nearly 2 million reads). The metagenome curves similarly approached saturation although both 2018 data sets would likely have benefited from deeper sequencing.

Eukaryotic taxonomic composition: averages across samples

Phylum Chlorophyta made up the largest portion of both the unassembled and assembled eukaryotic metagenome sequences on average (43% and 25%, respectively), followed by unclassified in reads (11%) and Bacillariophyta in contigs (19%). Haptophyta and Dinophyceae were the largest fractions of COI (26%) and 18S (27%) data sets, respectively (Fig. 2).

The Phylum Arthropoda was the second- and third-highest fraction of the 18S and COI amplicon data (23% and 18%, respectively), but made up only 2% and 3% of the metagenome reads and contigs, respectively (Fig. 3). Moreover, lower-level arthropod taxonomy differed substantially among the datasets. The largest fraction of arthropods in the metagenome reads were Malacostraca (64%) followed by Hexanauplia (15%) and Insecta (16%). Metagenome contigs were represented by Hexanauplia (44%) followed by Insecta (27%). COI sequences were mostly Hexanauplia (80%) followed by Malacostraca (12%)

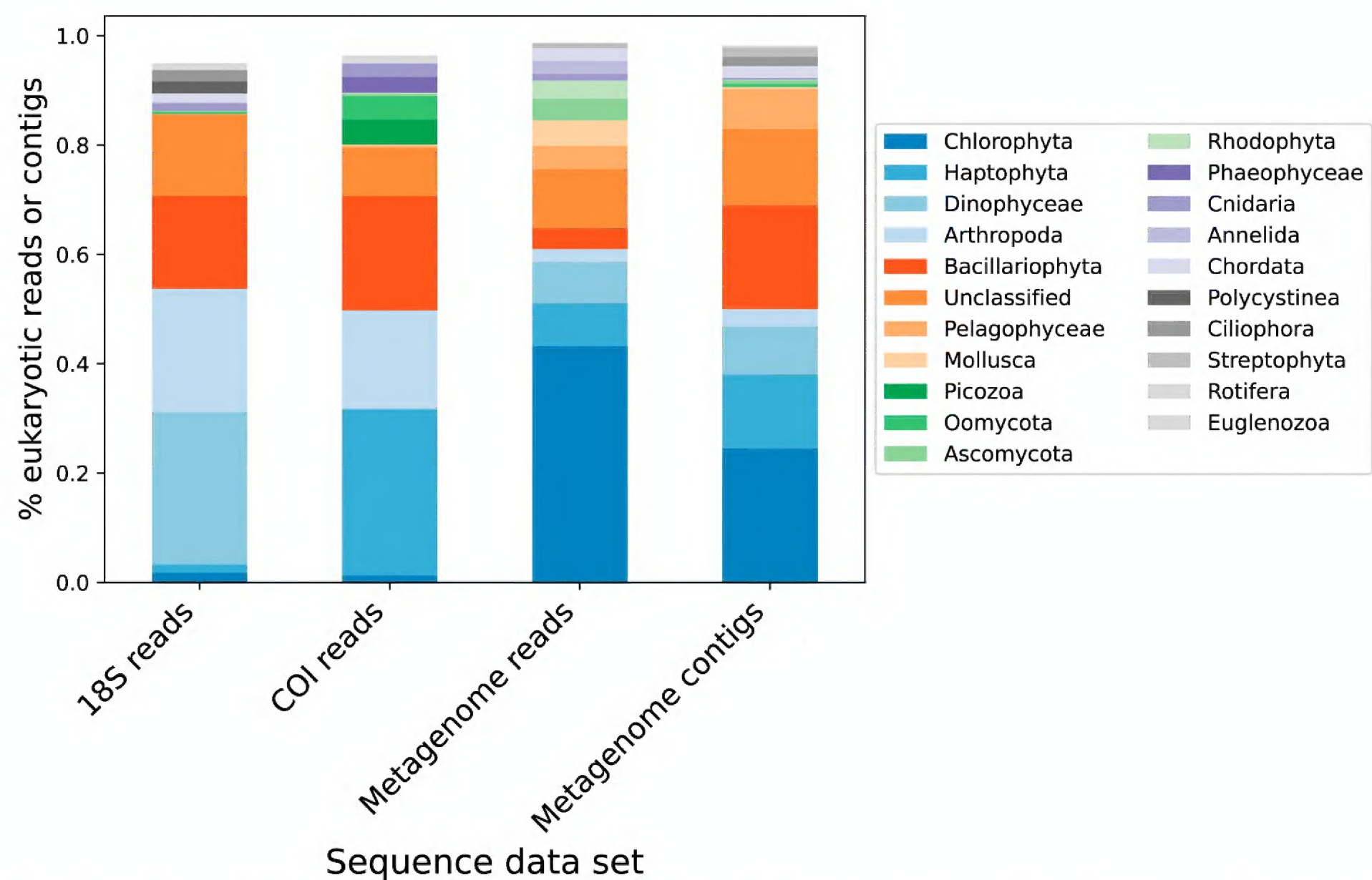


Figure 2. Average phylum-level observations across samples in each sequence data set. The phyla were limited to those with a maximum relative abundance greater than 1% in any of the four data sets.

and Insecta (5%), and 18S were mostly Hexanauplia (67%) and Malacostraca (29%), with everything else under 2%.

The Phylum Chordata, which includes all vertebrates, represented 2% of both the eukaryotic metagenome reads and contigs on average (Suppl. material 7: fig. S2). The metagenomes included Aves (birds), Actinopteri (teleost fish), and mammals (including human as well as *Megaptera novaeangliae*, or humpback whale). The 18S eukaryotic fraction also consisted of 2% Chordata, most of which were tunicates, salps, and ascidians. Chordata made up only 0.03% of the COI eukaryotic reads but included fishes and mammals, including 3 species of dolphins and *M. novaeangliae*. Other phyla with annotation discrepancies at lower taxonomic levels included Streptophyta and Mollusca.

Paired sample correlations

There were 38 (18S) and 41 (COI) phyla with a summed relative abundance across all samples greater than 1%. Three of the most prevalent phyla across all data sets (Chlorophyta, Haptophyta, and Bacillariophyta) had higher than average correlation coefficients among paired samples (Fig. 4, Fig. 5). Among these three phyla, the COI marker had comparable or higher correlations than 18S (Fig. 4). However, across all phyla, the 18S correlations were stronger on average and there were three times as many phyla with a coefficient greater than 0.4 compared to the COI correlations (9 vs 3; Fig. 5, Suppl. material 3: table S3). The highest correlation coefficient of all phyla was 0.75 for Discosea in the 18S data set (Suppl. material 3: table S3). However, Bacillariophyta was the only phylum

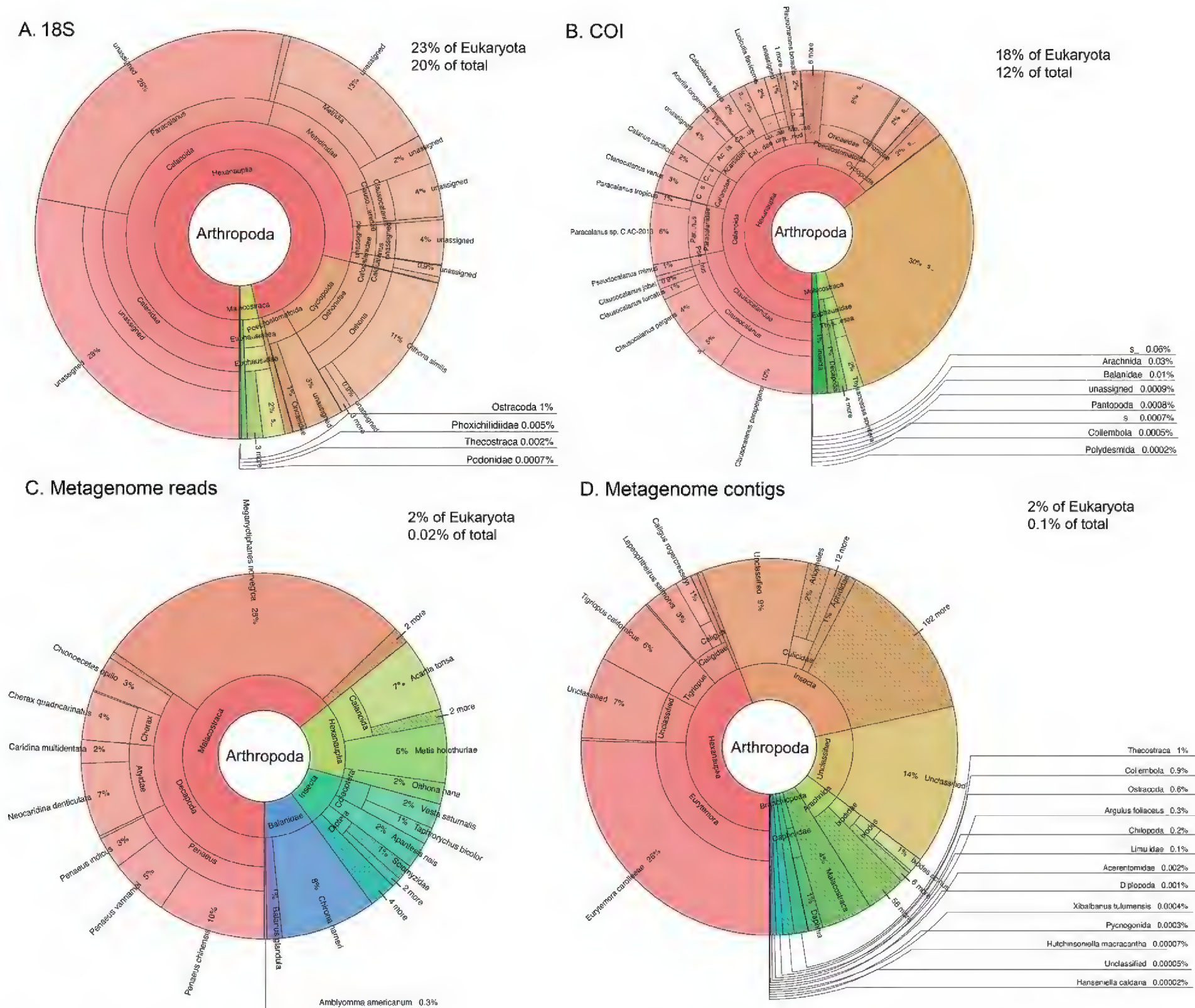


Figure 3. Krona plots for the Phylum Arthropoda showing the breakdown of taxonomic ranks in (A) 18S, (B) COI, (C) metagenome reads, and (D) metagenome contigs.

with a coefficient above 0.6 for both marker genes. The COI correlation coefficients for Phylum Dinophyceae were 0.15 for 18S and -0.15 for COI, reflecting the highly variable detections with marker genes (and overall poor performance of COI for dinoflagellates). One outlier COI sample contained a relative abundance of Dinophyceae ~0.15, represented entirely by one ASV comprising 4,498 out of a sample total of 30,561 reads (Suppl. material 8: fig. S3). This ASV was observed in two other samples but in each case with only a single read count. Phylogenetic analysis showed this ASV nested among all other COI Dinophyceae ASVs (Suppl. material 8: fig. S3), suggesting it was not mis-annotated but rather showed unusually high fidelity in amplification and library preparation.

Humpback whales (*Megaptera novaeangliae*) were detected in five COI samples, four of which had corresponding metagenomes with *M. novaeangliae* annotations.

The phylum-level correlations sometimes broke down at lower taxonomic levels. For instance, the Chlorophyta class Mamiellophyceae, which dominated the metagenome detections, was poor for 18S but relatively good for COI (Suppl. material 9: fig. S4). Class-level detections were also noticeably different among data sets for Bacillariophyta and Haptophyta (Suppl. material 9: fig. S4).

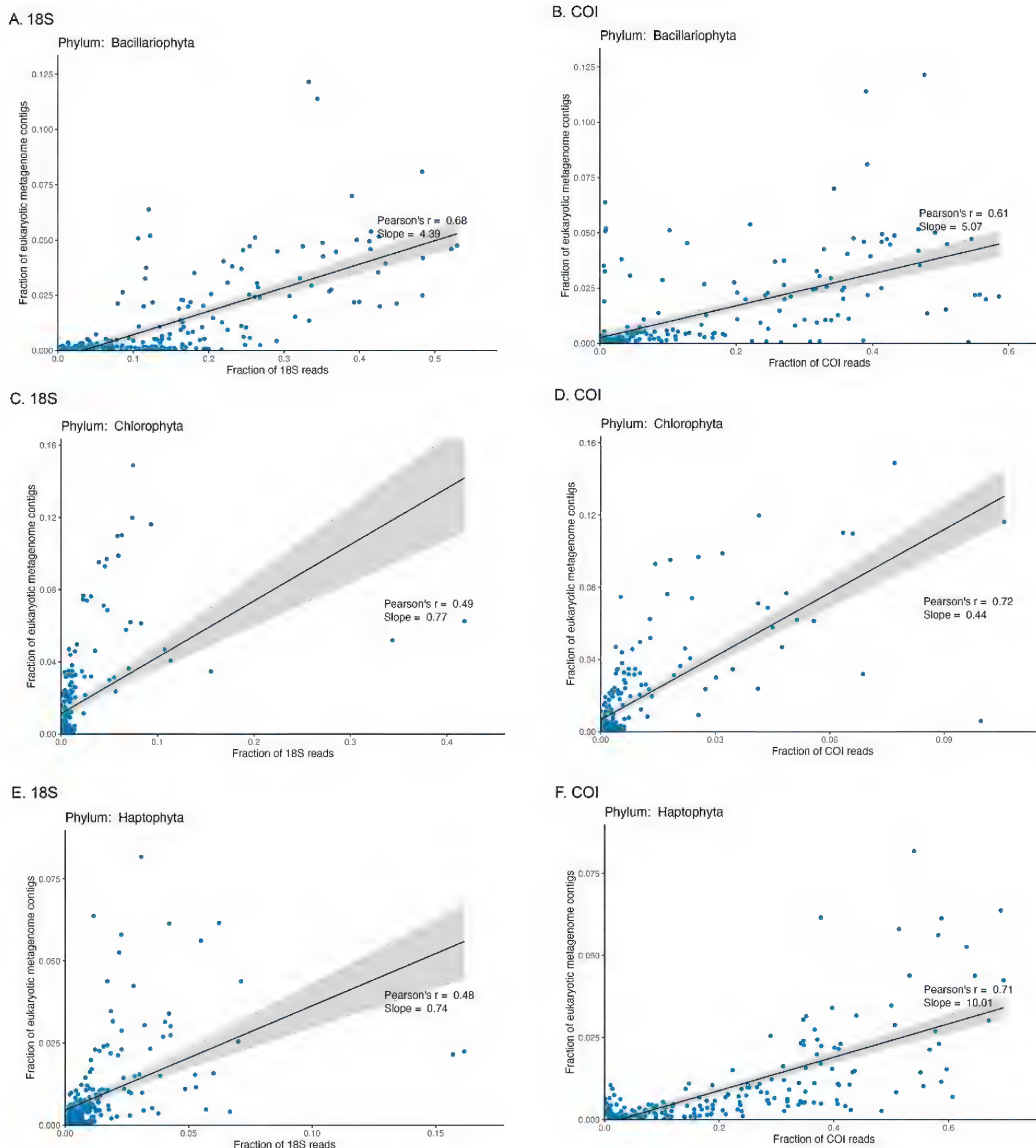


Figure 4. Six examples of phylum-level correlations among metagenomes and 18S (A, C, E) and COI (B, D, F) paired samples. Each point represents one eDNA sample with marker gene relative abundances on the x-axis and metagenome relative abundances for the same sample on the y-axis.

Presence of Streptophyta by depth and phylogeny of Magnoliopsida 18S ASVs

The Phylum Streptophyta, which was largely made up of flowering plant taxa like the family Magnoliopsida, was most prevalent in the integrated depth samples and had the lowest relative abundance in the surface water samples (Suppl. material 10: fig. S5).

Phylogenetic analysis of all 18S ASVs annotated as belonging to the Family Magnoliopsida (Kingdom Viridiplantae, Phylum Streptophyta) showed they

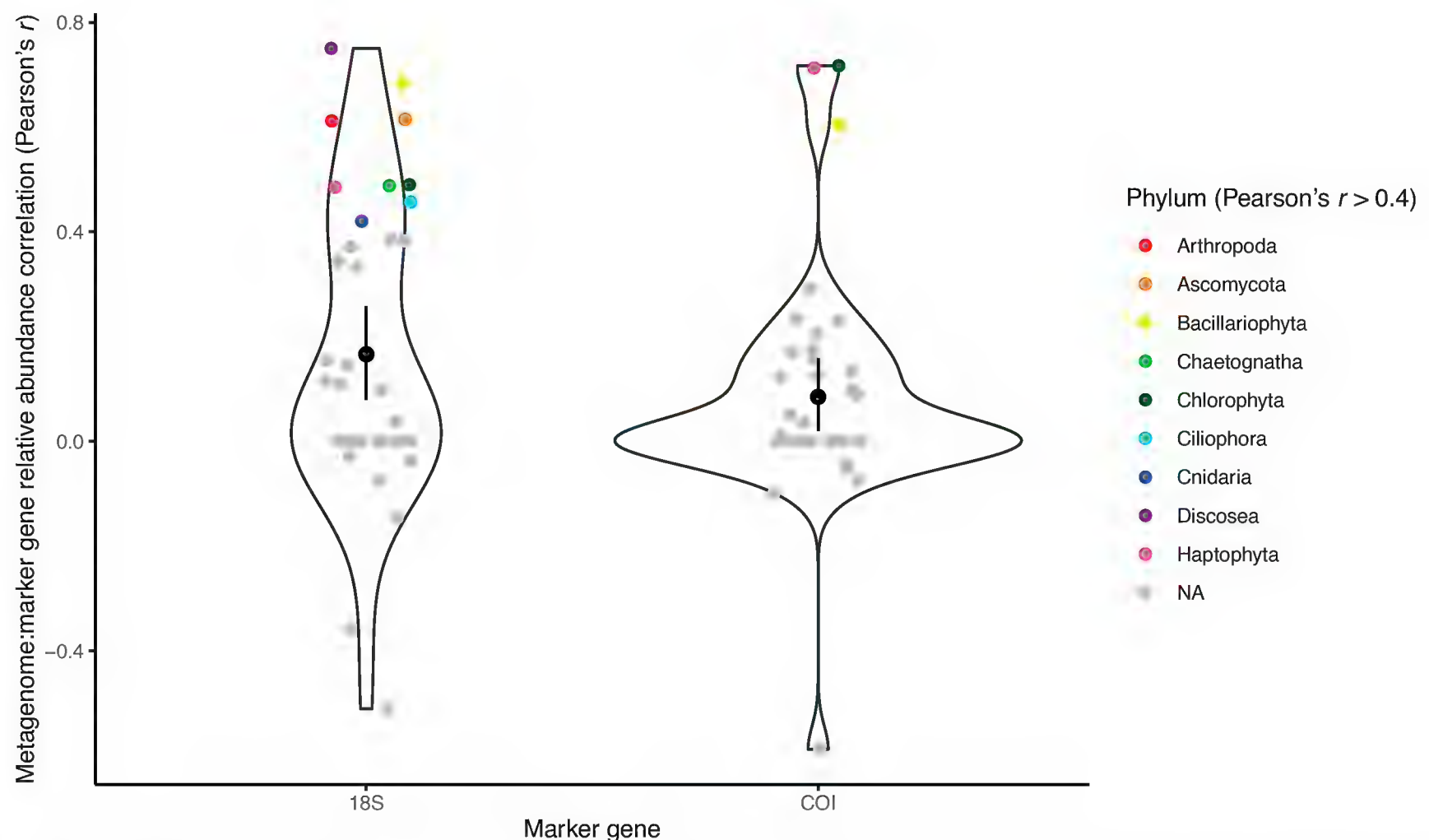


Figure 5. Phylum-level correlations between metagenome (assembled contigs) and amplicon relative abundances. Each point represents one phylum, for which a linear regression was calculated based on relative abundances in metagenome and amplicon sample pairs. Average correlation values for each data set are indicated by a black point and the vertical lines represent bootstrapped 95% confidence intervals. Phyla are labeled for which the correlation coefficient is greater than 0.4, indicating a strong correlation between relative abundances in metagenome and amplicon sample pairs.

were distantly related to reference sequences from that family (Suppl. material 11: fig. S6). Reference sequences from the Phyla Chlorophyta (Kingdom Viridiplantae) and Ascomycota (Kingdom Fungi) were more closely related to the Magnoliopsida reference sequences than the environmental ASVs.

Reference database composition of marker genes vs genomic nucleotides

There were notable differences among phyla in marker gene vs whole genome nucleotide composition of the reference database (NCBI). The Phyla Chordata, Streptophyta, and Phoronida had much higher nucleotide representation in whole genome sequences compared to both 18S rRNA and COI marker gene representation (Fig. 6). The Phylum Euglenozoa was also disproportionately high in whole genome nucleotides compared to COI genes while Echinodermata, Hemichordata, and Arthropoda had higher ratios relative to 18S rRNA marker genes. The following phyla had no whole genome sequences in GenBank: Prasinodermophyta, Phaeosacciophyceae, Picozoa, Raphidophyceae, Centroplasthelida, Dictyochophyceae, Chaetognatha, Acantharea, Gastrotricha, and Polycystinea (Suppl. material 5: table S5).

Discussion

Shotgun metagenomes are rarely used for eukaryotic surveys because detection of multicellular organisms is considered sparse and random (Takahashi et al. 2023). However, by minimizing the effects of PCR amplification, shotgun

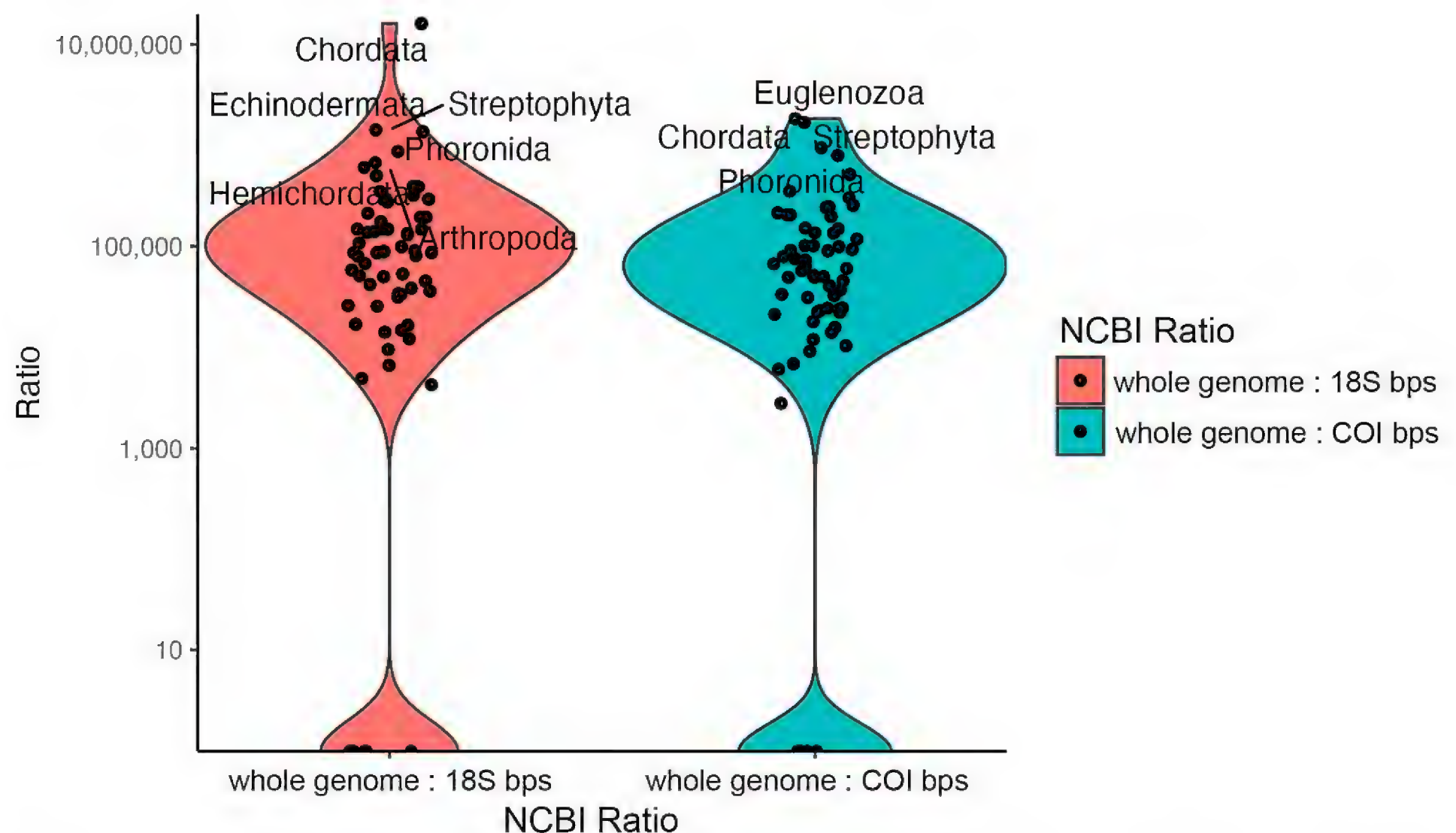


Figure 6. The ratio of whole genome to marker gene sequence content of NCBI for each phylum. Pyla with ratios above 6,000,000 are labeled.

metagenomes do not face the taxonomic detection limitations of marker gene surveys nor the skewed relative abundance values compared to starting values (Shelton et al. 2023). Here, we show that metagenomic data can be highly informative for eukaryotic taxonomic diversity, including metazoans. While markers designed for specific eukaryotic organisms, such as the 12S rRNA gene for fish, may be more effective for characterizing diversity of specific groups, here we address the interest in a microbes-to-mammals sequencing approach which is not possible with more targeted marker genes.

There is no way to know the “true” composition of eDNA from any one method, since each has its own biases. However, by comparing four approaches, we can assume that if results from two of them agree then those values are more likely to represent the original eDNA pool. Two of our data sets included PCR amplification of a marker gene while the shotgun metagenomic approaches did not. In theory, relative abundances from shotgun metagenomes should be more similar to “true” biological values because they were not affected by PCR primer and amplification biases. However, detection of eukaryotic taxa is likely influenced by the high levels of prokaryotic DNA, which may dominate some samples more than others and thus distort representation of more rare organisms. Rarefaction curves showed sequencing depth of most datasets likely encompassed a large portion of biological diversity, with the exception of the COI 2019 amplicon samples (Suppl. material 6: fig. S1). While taxonomic diversity of the amplicons was expected to be lower than metagenomes because they only captured the pool of marker genes rather than all possible nucleotide diversity, these samples were also sequenced an order of magnitude lower depth (Table 1); thus, it seems the applied sequencing effort was roughly appropriate for the diversity of the eDNA pool.

The differences between amplicon and metagenome annotations suggest some advantages to shotgun sequencing for assessing community diversity and evenness. Phyla observed at drastically different levels in the two amplicon data sets, such as Haptophyta (which include coccolithophores) and Dinophyceae (dinoflagellates), were both present at consistent levels in the metagenome reads and contigs (Fig. 2). These are two ecologically critical groups of phytoplankton that often dominate pelagic communities (Cohen et al. 2021; Vincent et al. 2022; Xu et al. 2022) and whose distributions are predicted to shift with global climate change (Benedetti et al. 2021); more accurate assessments of their populations would be highly valuable to improving our understanding of phytoplankton contributions to global nutrient cycles, and shotgun metagenomics may offer an advantage in that regard. Dinoflagellates in particular are known to amplify poorly with the COI primers (Lin et al. 2009; Min et al. 2023) and to have many copies of rRNA genes (Zhu et al. 2005); they are therefore likely to be under-represented by COI and over-represented by 18S rRNA marker gene sequencing. Metagenomes also featured relatively high levels of phyla that were nearly absent from both amplicon data sets. These included Chlorophyta and Pelagophyceae, two ubiquitous groups of microalgae whose ecology is poorly understood but whose contributions to photosynthesis may be significant (Vaulot et al. 2008; Worden et al. 2012; Tragin and Vaulot 2019). Both groups have been explored more recently with the advent of molecular methods, and shotgun metagenomes may provide a better overview of their abundance and diversity relative to marker gene sequencing based on our data. Conversely, both amplicon data sets detected relatively high levels of Arthropoda, which composed <1% of the metagenomes). Taxa like calanoid copepods and euphausiid krill, both of which are highly abundant in the CCE (Lavaniegos and Ohman 2007), were identified to the species level by the marker genes while metagenomes featured larger proportions of insects and arachnids. Many of the genus- and species-level annotations in metagenomes were also clearly incorrect, e.g. freshwater crayfish and North Sea copepods. Thus, while phylum-level proportions may have been more accurately represented in the metagenomes, the species-level identifications were more reliable in the marker gene data sets and would be more appropriate for zooplankton community assessment.

Diatoms (Phylum Bacillariophyta) are another ecologically critical group of phytoplankton and were observed at similar levels both on average across samples and between paired samples (Fig. 4, Suppl. material 3: table S3). Composition at lower taxonomic levels was also similar for the diatoms, with Coscinodiphyceae and Bacillariophyceae comprising the bulk of both metagenome and amplicon sequences (Suppl. material 9: fig. S4). However, the agreement between metagenomes and marker genomes for this phylum was an anomaly, which (apart from the “Unclassified” group) was rarely observed at similar levels both on average across samples (Fig. 2) or between matched samples (Fig. 4).

One potential application of shotgun metagenomics is for detection of multiple organisms which would otherwise require several different marker genes or single species assays. The Phylum Chordata encompasses many organisms that may be of conservation and management interest, such as teleost fish. While all data sets featured low levels of chordates, only the COI and metagenomic data identified important fish and cetacean species (which showed notable agreement on

a per-sample basis) and only metagenomes had annotations for the Class Aves (birds). Considering the different scales of sequencing depth between amplicon and metagenomic data, shotgun metagenomes certainly provided the most sequence data for this group at the Phylum level. However, the poor accuracy of the metagenomic species-level annotations is problematic for an accurate assessment of chordate diversity (Suppl. material 7: fig. S2). Both metagenome data sets showed a tunicate species in the genus *Oikopleura* making up about one-third of the chordate annotations but differed in the species annotation. They also featured several bird, fish, and mammal species that do not occur in the CCE. While some species like humans, dogs, and cows may constitute lab contamination that could be accounted for, others like bats and camels would require extensive manual analysis to annotate correctly. In contrast, the COI marker identified three dolphin species known to inhabit the study area, and did not have any likely mammalian contamination other than human DNA. Although chordate genomes are well represented in NCBI relative to COI genes (Fig. 6), the diversity of functional genes and their wide range of evolutionary rates relative to a marker like COI likely contribute to the likely inaccurate annotations of metagenomic contigs. In general, clade-specific marker genes are likely better suited for rare eukaryotic targets than either general markers like the two used in this study or shotgun metagenomes.

The differences in whole genome and marker gene representation in the reference database (NCBI / UniRef) affected the annotations within certain phyla. For instance, Arthropoda and Streptophyta both featured high levels of terrestrial organisms (insects and arachnids for arthropods and flowering plants for streptophytes). While terrestrial transport of flowering plant biomass could affect coastal samples, these annotations were at their lowest level in the surface samples (Suppl. material 10: fig. S5) where terrestrial influence would be highest. This discrepancy thus likely reflects the mis-annotation of planktonic taxa with little to no representation in NCBI. Arthropoda, Streptophyta, and Chordata were all disproportionately represented by whole genome nucleotides in NCBI relative to marker genes (Fig. 6); thus, there is much greater sequence space available for alignment and annotation for a clade lacking any database representation compared to other taxonomic groups. Picoeukaryotic phytoplankton have historically been neglected due to challenges of cultivation and high levels of genetic and functional diversity (Decelle et al. 2015; Caron et al. 2017; Keeling and Campo 2017), and it is possible there are photosynthetic clades highly represented in the eDNA pool both lacking genomic representation in NCBI and whose marker genes are not well amplified by the 18S and COI primers. Some 18S amplicon samples showed low levels of “Magnoliopsida,” the flowering plant class detected in the metagenomes; however, all ASVs annotated as such showed a distant phylogenetic relationship to reference genes from three species in the family Magnoliopsida (Suppl. material 11: fig. S6), suggesting they belong to a clade that lacks representation in NCBI. Phoronida, Euglenozoa, and Echinodermata also had disproportionately higher levels of representation but these phyla were not detected above 1% on average across all samples within any data set; thus, the differences in database composition were less impactful for the results.

Other factors may have affected the observed differences in annotations. For instance, we used whole genome sequence information for the comparison of database representation because annotations of individual genes are much more challenging to quantify and verify; however, there may be high nucleotide

content of genes not contained in whole genome sequences that would impact annotation levels and would not have been captured in our comparison. Metagenomic data are also affected by factors like genome size and, for multicellular organisms, the number of cells captured from a single organism, which can skew the relative contributions of eDNA. Moreover, if many of the annotated genes belonged to the universal single-copy orthologous groups (BUSCO; Simão et al. 2015), they could easily have been mis-annotated because orthologs are highly conserved across the tree of life. For example, if the samples contained high levels of DNA from single-celled picoeukaryotic organisms with no marker genes or whole genomes in NCBI, their BUSCO genes may have been annotated to a relatively distantly related clade, such as the terrestrial flowering plant family Magnoliopsida, due to the Least Common Ancestor approach for annotation.

Several factors other than primer selectivity may also affect the differences in correlations with metagenomic data between the two marker genes. First, the COI gene is mitochondrial while the 18S rRNA gene is nuclear; there may be quantitative implications to the different organellar sources. McCauley et al. 2024 showed that nuclear DNA persists longer in the environment, likely due to its more resilient membrane. Moreover, different rates of evolution and inter-specific diversity may necessitate different bioinformatic approaches between the two genes (Antich et al. 2021). While curated marker gene-specific databases (e.g. MetaZooGene, PR2) are likely better for annotation of specific organism groups, they do not offer the full taxonomic breadth of NCBI. Additionally, we performed direct sample pairwise comparisons with values calculated from short reads mapped to assembled contigs, which only encompass a portion of the metagenome due to computational challenges with assembly (Nurk et al. 2017; Nayfach et al. 2019; Olson et al. 2019; Saraiva et al. 2023). On average, between 30 and 40% of short reads mapped back to the contigs, suggesting that unassembled reads may offer additional diversity not captured in our analysis. However, longer contigs annotated with LCA should provide more accurate annotations relative to short reads, and we expect most researchers would prefer to use assemblies for estimating taxonomic composition from shotgun metagenomes.

Marker genes are highly useful for selecting against prokaryotic DNA, which dominates marine environmental samples. This advantage remains critical for maximizing sequencing efficiency and identifying species-level diversity in eukaryotes, particularly as reference databases remain deficient in whole genomes across large swaths of the Tree of Life. However, primer and amplification biases clearly prevent detection of critical taxa in their natural proportions. Shotgun metagenomics can, contrary to common dogma, detect even the rarest metazoan taxa like cetaceans. As reference databases continue to become populated with whole genome sequences and annotation methods (particularly k-mer based approaches) continue to improve, the microbes-to-mammals molecular surveys will become more tractable and reduce the need for resource-intensive and highly biased amplifications.

Conclusions

Our direct comparison of amplicon and metagenome sequencing of >200 marine water column samples reveal the qualitative advantages and drawbacks of each method for assessing marine eukaryotic communities. While marker gene

sequencing can provide fine-scale taxonomic resolution of particular clades of interest and relative abundances of taxa with relatively little computational effort, the massive scale of shotgun metagenomics overcomes the disproportionate prokaryotic representation to the extent that eukaryotic sequence levels are still two orders magnitude higher in metagenomes. The amount of information that can be extracted from these data is still limited by the ability to accurately assemble eukaryotic genes, as the assembly process leads to an overall drop in sequence space for all represented domains. Moreover, while avoiding PCR can prevent biases in phylum-level quantitative assessments, species- and genus-level taxonomic annotations are frequently inaccurate. These misleading results are due at least in part to reference database limitations, in particular the under-representation of marine taxa in whole genome sequence data. Generating custom databases limited to taxa known to occur in the sampled habitat may provide more confident annotations; however, only additional sequencing of under-studied taxa, such as eukaryotic protists and invertebrates, will provide a universal solution.

Shotgun metagenomes undeniably have a higher per-sample cost relative to metabarcoding. However, the value demonstrated here should incentivize metagenomic sequencing where possible. One potential option is to replace the PhiX genome, used as a standard spike for Illumina amplicon sequencing to increase nucleotide diversity in the library, with one or more metagenomes. The nucleotide diversity should offer the same benefit to the sequencing chemistry as PhiX and would provide additional valuable data, rather than using up to 20% of the sequencing run for generating data that will be discarded. Moreover, shotgun metagenomes can act as a molecular archive, which can be re-interrogated with improved bioinformatic methods as they arise.

Both metabarcoding and shotgun metagenomes can provide valuable biodiversity information for eukaryotic organisms, and often the information is complementary. As sequencing costs continue to decrease and reference databases improve, we anticipate greater application of metagenomics and hope that the information provided here will assist researchers in experimental design and lead to more accurate and informative biodiversity data.

Acknowledgements

We thank the NOAA Fisheries scientists of the Rockfish Recruitment and Ecosystem Assessment Survey, in particular John Field and Keith Sakuma, and the crew of the NOAA Ship Reuben Lasker for facilitating eDNA sampling. We acknowledge the High Performance Computing resources provided by Mississippi State University and the Northern Gulf Institute, which supported the computational analyses. We also gratefully acknowledge our NOAA 'Omics colleagues for thoughtful discussions and suggestions regarding data analysis. The generous support of the David and Lucile Packard Foundation to MBARI is also acknowledged.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Use of AI

No use of AI was reported.

Funding

This work was supported by the U.S. Marine Biodiversity Observation Network (MBON) co-organized by NOAA, NASA, BOEM, and ONR through the National Oceanographic Partnership Program (NOPP) via NASA award 80NSSC21M0032 and NOAA award NA22NOS0120184 to FPC. It was carried out in part under the auspices of the Cooperative Institute of Marine and Atmospheric Studies (CIMAS), a Cooperative Institute of the University of Miami and NOAA, under cooperative agreement number NA20OAR4320472. We also acknowledge the NOAA/OAR 'Omics Program (NOAA 'Omics OAR NO_0049).

Author contributions

NVP conceived of the study, collected samples, performed the analyses, and wrote the manuscript. KP collected samples, performed analyses and edited the manuscript. JB processed samples and edited the manuscript. FC and KDG supported the study at all stages.

Author ORCIDs

Nastassia V. Patin  <https://orcid.org/0000-0001-8522-7682>

Kathleen J. Pitz  <https://orcid.org/0000-0002-4931-8592>

Jacoby D. Baker  <https://orcid.org/0000-0002-0673-7535>

Francisco P. Chavez  <https://orcid.org/0000-0002-0691-292X>

Kelly D. Goodwin  <https://orcid.org/0000-0001-9583-8073>

Data availability

All custom code referenced in this manuscript is available on GitHub at <https://github.com/nvpatin/Euk-eDNA-comparison>. All phylum- and class-level correlation plots are on FigShare (18S phylum-level: <https://doi.org/10.6084/m9.figshare.26065150.v2>, 18S class-level: <https://doi.org/10.6084/m9.figshare.26304682.v1>, COI phylum-level: <https://doi.org/10.6084/m9.figshare.26065141.v1>, COI class-level: <https://doi.org/10.6084/m9.figshare.26304700>). Sequence data is available on NCBI Sequence Read Archive under BioProjects PRJNA853328 (metagenomes) and PRJNA726280 (amplicons).

References

- Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution* 9: 134–147. <https://doi.org/10.1111/2041-210X.12849>
- Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A method for studying Protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* 4: e6372. <https://doi.org/10.1371/journal.pone.0006372>
- Andres KJ, Lodge DM, Andrés J (2023) Environmental DNA reveals the genetic diversity and population structure of an invasive species in the Laurentian Great Lakes. *Proceedings of the National Academy of Sciences of the United States of America* 120: e2307345120. <https://doi.org/10.1073/pnas.2307345120>

- BMTagger (2022) BMTagger. [Available from:] <ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/> [Accessed February 16, 2022]
- Antich A, Palacin C, Wangenstein OS, Turon X (2021) To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics* 22: 177. <https://doi.org/10.1186/s12859-021-04115-6>
- Apprill A, McNally S, Parsons R, Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* 75: 129–137. <https://doi.org/10.3354/ame01753>
- Arita M, Karsch-Mizrachi I, Cochrane G (2021) The international nucleotide sequence database collaboration. *Nucleic Acids Research* 49: D121–D124. <https://doi.org/10.1093/nar/gkaa967>
- Aroney STN, Newell RJP, Nissen JN, Camargo AP, Tyson GW, Woodcroft BJ (2025) CoverM: read alignment statistics for metagenomics. *Bioinformatics* 41: btaf147. <https://doi.org/10.1093/bioinformatics/btaf147>
- Benedetti F, Vogt M, Elizondo UH, Righetti D, Zimmermann NE, Gruber N (2021) Major restructuring of marine plankton assemblages under global warming. *Nature Communications* 12: 5226. <https://doi.org/10.1038/s41467-021-25385-x>
- Bianchini G, Sánchez-Baracaldo P (2024) TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecology and Evolution* 14: e10873. <https://doi.org/10.1002/ece3.10873>
- Bucklin A, Peijnenburg KTCA, Kosobokova KN, O'Brien TD, Blanco-Bercial L, Cornils A, Falkenhaus T, Hopcroft RR, Hosia A, Laakmann S, Li C, Martell L, Questel JM, Wall-Palmer D, Wang M, Wiebe PH, Weydmann-Zwolicka A (2021) Toward a global reference database of COI barcodes for marine zooplankton. *Marine Biology* 168: 78. <https://doi.org/10.1007/s00227-021-03887-y>
- Bylemans J, Gleeson DM, Hardy CM, Furlan E (2018) Toward an ecoregion scale evaluation of EDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray–Darling Basin (Australia). *Ecology and Evolution* 8: 8697–8712. <https://doi.org/10.1002/ece3.4387>
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13: 581–583. <https://doi.org/10.1038/nmeth.3869>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421. <https://doi.org/10.1186/1471-2105-10-421>
- Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, Bell CJ, Bharti A, Dyhrman ST, Guida SM, Heidelberg KB, Kaye JZ, Metzner J, Smith SR, Worden AZ (2017) Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology* 15: 6–20. <https://doi.org/10.1038/nrmicro.2016.160>
- Casey JM, Ransome E, Collins AG, Mahardini A, Kurniasih EM, Sembiring A, Schiettekatte NMD, Cahyani NKD, Wahyu Anggoro A, Moore M, Uehling A, Belcaid M, Barber PH, Geller JB, Meyer CP (2021) DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. *Environmental DNA* 3: 1229–1246. <https://doi.org/10.1002/edn3.245>
- Chamberlain S, Arendsee Z, Stirling T, Boettiger C, James TD, Salmon M, Li G, Grenié M (2023) ropensci/taxizedb: taxizedb v0.3.1. <https://doi.org/10.5281/ZENODO.1158055>
- Chen S (2023) Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* 2: e107. <https://doi.org/10.1002/imt2.107>

- Chou S, Reiter T (2024) A new R package, sourmashconsumr, for analyzing and visualizing the outputs of sourmash. Software: Implementing useful and innovative computing. <https://doi.org/10.57844/ARCADIA-1896-KE33>
- Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution* 7: 873–883. <https://doi.org/10.1002/ece3.2667>
- Closek CJ, Santora JA, Starks HA, Schroeder ID, Andruszkiewicz EA, Sakuma KM, Bograd SJ, Hazen EL, Field JC, Boehm AB (2019) Marine vertebrate biodiversity and distribution within the Central California current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. *Frontiers in Marine Science* 6: 732. <https://doi.org/10.3389/fmars.2019.00732>
- Cohen NR, McIlvin MR, Moran DM, Held NA, Saunders JK, Hawco NJ, Brosnahan M, DiTullio GR, Lamborg C, McCrow JP, Dupont CL, Allen AE, Saito MA (2021) Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology* 6: 173–186. <https://doi.org/10.1038/s41564-020-00814-7>
- Collins RA, Bakker J, Wangenstein OS, Soto AZ, Corrigan L, Sims DW, Genner MJ, Mariani S (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution* 10: 1985–2001. <https://doi.org/10.1111/2041-210X.13276>
- Compson ZG, McClenaghan B, Singer GAC, Fahner NA, Hajibabaei M (2020) Metabarcoding from microbes to mammals: Comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution* 8: 581835. <https://doi.org/10.3389/fevo.2020.581835>
- Cowart DA, Pinheiro M, Mouchel O, Maguer M, Grall J, Miné J, Arnaud-Haond S (2015) Metabarcoding Is powerful yet still blind: A comparative analysis of morphological and molecular surveys of seagrass communities. *PLOS ONE* 10: e0117562. <https://doi.org/10.1371/journal.pone.0117562>
- Cowart DA, Murphy KR, Cheng C-HC (2018) Metagenomic sequencing of environmental DNA reveals marine faunal assemblages from the West Antarctic Peninsula. *Marine Genomics* 37: 148–160. <https://doi.org/10.1016/j.margen.2017.11.003>
- Curd EE, Gal L, Gallego R, Silliman K, Nielsen S, Gold Z (2024) rCRUX: A rapid and versatile tool for generating metabarcoding reference libraries in R. *Environmental DNA* 6: e489. <https://doi.org/10.1002/edn3.489>
- De Sena Brandine G, Smith AD (2021) Falco: High-speed FastQC emulation for quality control of sequencing data. *F1000 Research* 8: 1874. <https://doi.org/10.12688/f1000research.21142.2>
- Decelle J, Colin S, Foster RA (2015) Photosymbiosis in Marine Planktonic Protists. In: Ohtsuka S, Suzuki T, Horiguchi T, Suzuki N, Not F (Eds) *Marine Protists*. Springer Japan, Tokyo, 465–500. https://doi.org/10.1007/978-4-431-55130-0_19
- Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, De Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology* 26: 5872–5895. <https://doi.org/10.1111/mec.14350>
- Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, Eren AM, Kourlaiev A, d'Agata L, Clayssen Q, Villar E, Labadie K, Cruaud C, Poulain J, Da Silva C, Wessner M, Noel B, Aury J-M, De Vargas C, Bowler C, Karsenti E, Pelletier E, Wincker P, Jaillon O, Sunagawa S, Acinas SG, Bork P, Karsenti E, Bowler C, Sardet C, Stemmann L, De Vargas C, Wincker P, Lescot M, Babin M, Gorsky G, Grimsley N, Guidi L,

- Hingamp P, Jaillon O, Kandels S, Iudicone D, Ogata H, Pesant S, Sullivan MB, Not F, Lee K-B, Boss E, Cochrane G, Follows M, Poulton N, Raes J, Sieracki M, Speich S (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 2: 100123. <https://doi.org/10.1016/j.xgen.2022.100123>
- Didion JP, Martin M, Collins FS (2017) Atropos: Specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* 5: e3720. <https://doi.org/10.7717/peerj.3720>
- Drost H-G, Paszkowski J (2017) Biomart: genomic data retrieval with R. *Bioinformatics* 33: 1216–1217. <https://doi.org/10.1093/bioinformatics/btw821>
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3: 294–299.
- Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boute C, Burgaud G, De Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaultot D, Zimmermann P, Christen R (2012) The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41: D597–D604. <https://doi.org/10.1093/nar/gks1160>
- Harrison JB, Sunday JM, Rogers SM (2019) Predicting the fate of eDNA in the environment and implications for studying biodiversity. *Proceedings of the Royal Society B, Biological Sciences* 286: 20191409. <https://doi.org/10.1098/rspb.2019.1409>
- Huson DH, Beier S, Flade I, Górski A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R (2016) MEGAN Community Edition - Interactive exploration and analysis of large-scale microbiome sequencing data. *PLOS Computational Biology* 12: e1004957. <https://doi.org/10.1371/journal.pcbi.1004957>
- Keck F, Couton M, Altermatt F (2023) Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources* 23: 742–755. <https://doi.org/10.1111/1755-0998.13746>
- Keeling PJ, Campo JD (2017) Marine protists are not just big bacteria. *Current Biology* 27: R541–R549. <https://doi.org/10.1016/j.cub.2017.03.075>
- Kelly RP, Shelton AO, Gallego R (2019) Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports* 9: 12133. <https://doi.org/10.1038/s41598-019-48546-x>
- Larsson A (2014) AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics (Oxford, England)* 30: 3276–3278. <https://doi.org/10.1093/bioinformatics/btu531>
- Lavaniegos BE, Ohman MD (2007) Coherence of long-term variations of zooplankton in two sectors of the California Current System. *Progress in Oceanography* 75: 42–69. <https://doi.org/10.1016/j.pocean.2007.07.002>
- Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology* 10: 34. <https://doi.org/10.1186/1742-9994-10-34>
- Leray M, Knowlton N, Ho S-L, Nguyen BN, Machida RJ (2019) GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America* 116: 22651–22656. <https://doi.org/10.1073/pnas.1911714116>

- Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* (Oxford, England) 26: 493–500. <https://doi.org/10.1093/bioinformatics/btp692>
- Lin S, Zhang H, Hou Y, Zhuang Y, Miranda L (2009) High-level diversity of Dinoflagellates in the natural environment, revealed by assessment of mitochondrial *cox1* and *cob* genes for Dinoflagellate DNA barcoding. *Applied and Environmental Microbiology* 75: 1279–1290. <https://doi.org/10.1128/AEM.01578-08>
- Lo C-C, Chain PSG (2014) Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics* 15: 366. <https://doi.org/10.1186/s12859-014-0366-2>
- Macher T-H, Schütz R, Yildiz A, Beermann AJ, Leese F (2023) Evaluating five primer pairs for environmental DNA metabarcoding of Central European fish species based on mock communities. *Metabarcoding and Metagenomics* 7: e103856. <https://doi.org/10.3897/mbmg.7.103856>
- Machida RJ, Kurihara H, Nakajima R, Sakamaki T, Lin Y-Y, Furusawa K (2021) Comparative analysis of zooplankton diversities and compositions estimated from complement DNA and genomic DNA amplicons, metatranscriptomics, and morphological identifications. *ICES Journal of Marine Science* 78: 3428–3443. <https://doi.org/10.1093/icesjms/fsab084>
- McCauley M, Koda SA, Loesgen S, Duffy DJ (2024) Multicellular species environmental DNA (eDNA) research constrained by overfocus on mitochondrial DNA. *The Science of the Total Environment* 912: 169550. <https://doi.org/10.1016/j.scitotenv.2023.169550>
- Min MA, Needham DM, Sudek S, Truelove NK, Pitz KJ, Chavez GM, Poirier C, Gardeler B, Von Der Esch E, Ludwig A, Riebesell U, Worden AZ, Chavez FP (2023) Ecological divergence of a mesocosm in an eastern boundary upwelling system assessed with multi-marker environmental DNA metabarcoding. *Biogeosciences* 20: 1277–1298. <https://doi.org/10.5194/bg-20-1277-2023>
- Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E (2021) Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 37: 3029–3031. <https://doi.org/10.1093/bioinformatics/btab184>
- Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC (2019) New insights from uncultivated genomes of the global human gut microbiome. *Nature* 568: 505–510. <https://doi.org/10.1038/s41586-019-1058-x>
- NCBI Resource Coordinators, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, Bolton E, Bourexis D, Brister JR, Bryant SH, Canese K, Cavanaugh M, Charowhas C, Clark K, Dondoshansky I, Feolo M, Fitzpatrick L, Funk K, Geer LY, Gorelenkov V, Graeff A, Hlavina W, Holmes B, Johnson M, Kattman B, Khotomlianski V, Kimchi A, Kimelman M, Kimura M, Kitts P, Klimke W, Kotliarov A, Krasnov S, Kuznetsov A, Landrum MJ, Landsman D, Lathrop S, Lee JM, Leubsdorf C, Lu Z, Madden TL, Marchler-Bauer A, Malheiro A, Meric P, Karsch-Mizrachi I, Mnev A, Murphy T, Orris R, Ostell J, O’Sullivan C, Palanigobu V, Panchenko AR, Phan L, Pierov B, Pruitt KD, Rodarmer K, Sayers EW, Schneider V, Schoch CL, Schuler GD, Sherry ST, Siyan K, Soboleva A, Soussov V, Starchenko G, Tatusova TA, Thibaud-Nissen F, Todorov K, Trawick BW, Vakarov D, Ward M, Yaschenko E, Zasytkin A, Zbicz K (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 46: D8–D13. <https://doi.org/10.1093/nar/gkx1095>

- Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: A new versatile metagenomic assembler. *Genome Research* 27: 824–834. <https://doi.org/10.1101/gr.213959.116>
- Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, Pop M (2019) Metagenomic assembly through the lens of validation: Recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics* 20: 1140–1150. <https://doi.org/10.1093/bib/bbx098>
- Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* 12: 385. <https://doi.org/10.1186/1471-2105-12-385>
- Patin NV, Goodwin KD (2022) Long-read sequencing improves recovery of Picoeukaryotic genomes and zooplankton marker genes from marine metagenomes. *mSystems* 7: e00595-22. <https://doi.org/10.1128/msystems.00595-22>
- Pierce NT, Irber L, Reiter T, Brooks P, Brown CT (2019) Large-scale sequence comparisons with sourmash. *F1000 Research* 8: 1006. <https://doi.org/10.12688/f1000research.19675.1>
- Piñol J, Mir G, Gomez-Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources* 15: 819–830. <https://doi.org/10.1111/1755-0998.12355>
- Pont D, Meulenbroek P, Bammer V, Dejean T, Erős T, Jean P, Lenhardt M, Nagel C, Pekarik L, Schabuss M, Stoeckle BC, Stoica E, Zornig H, Weigand A, Valentini A (2023) Quantitative monitoring of diverse fish communities on a large scale combining eDNA metabarcoding and qPCR. *Molecular Ecology Resources* 23: 396–409. <https://doi.org/10.1111/1755-0998.13715>
- Renner SS, Scherz MD, Schoch CL, Gottschling M, Vences M (2024) Improving the gold standard in NCBI GenBank and related databases: DNA sequences from type specimens and type strains. *Systematic Biology* 73: 486–494. <https://doi.org/10.1093/sysbio/syad068>
- Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research* 39: e145–e145. <https://doi.org/10.1093/nar/gkr732>
- Ruppert KM, Kline RJ, Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* 17: e00547. <https://doi.org/10.1016/j.gecco.2019.e00547>
- Saraiva JP, Bartholomäus A, Toscan RB, Baldrian P, Nunes da Rocha U (2023) Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes. *Molecular Ecology Resources* 23: 1066–1076. <https://doi.org/10.1111/1755-0998.13776>
- Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Farrell CM, Feldgarden M, Fine AM, Funk K, Hatcher E, Kannan S, Kelly C, Kim S, Klimke W, Landrum MJ, Lathrop S, Lu Z, Madden TL, Malheiro A, Marchler-Bauer A, Murphy TD, Phan L, Pujar S, Rangwala SH, Schneider VA, Tse T, Wang J, Ye J, Trawick BW, Pruitt KD, Sherry ST (2023) Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Research* 51: D29–D38. <https://doi.org/10.1093/nar/gkac1032>

- Serite CP, Emami-Khoyi A, Ntshudisane OK, James NC, Jansen Van Vuuren B, Bodill T, Cowley PD, Whitfield AK, Teske PR (2023) eDNA metabarcoding vs metagenomics: An assessment of dietary competition in two estuarine pipefishes. *Frontiers in Marine Science* 10: 1116741. <https://doi.org/10.3389/fmars.2023.1116741>
- Shelton AO, Gold ZJ, Jensen AJ, D'Agnese E, Andruszkiewicz Allan E, Van Cise A, Gallego R, Ramón-Laca A, Garber-Yonts M, Parsons K, Kelly RP (2023) Toward quantitative metabarcoding. *Ecology* 104: e3906. <https://doi.org/10.1002/ecy.3906>
- Shen W, Le S, Li Y, Hu F (2016) SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLOS ONE* 11: e0163962. <https://doi.org/10.1371/journal.pone.0163962>
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* 31: 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Singer GAC, Shekarriz S, McCarthy A, Fahner N, Hajibabaei M (2020) The utility of a metagenomics approach for marine biomonitoring. *bioRxiv*. <https://doi.org/10.1101/2020.03.16.993667>
- Srivathsan A, Sha JCM, Vogler AP, Meier R (2015) Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources* 15: 250–261. <https://doi.org/10.1111/1755-0998.12302>
- Stat M, Huggett MJ, Bernasconi R, DiBattista JD, Berry TE, Newman SJ, Harvey ES, Bunce M (2017) Ecosystem biomonitoring with eDNA: Metabarcoding across the tree of life in a tropical marine environment. *Scientific Reports* 7: 12240. <https://doi.org/10.1038/s41598-017-12501-5>
- Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology* 19: 21–31. <https://doi.org/10.1111/j.1365-294X.2009.04480.x>
- Strasser BJ (2008) GenBank—Natural History in the 21st Century? *Science* 322: 537–538. <https://doi.org/10.1126/science.1163399>
- Takahashi M, Saccò M, Kestel JH, Nester G, Campbell MA, Van Der Heyde M, Heydenrych MJ, Juszkievicz DJ, Nevill P, Dawkins KL, Bessey C, Fernandes K, Miller H, Power M, Mousavi-Derazmahalleh M, Newton JP, White NE, Richards ZT, Allentoft ME (2023) Aquatic environmental DNA: A review of the macro-organismal biomonitoring revolution. *The Science of the Total Environment* 873: 162322. <https://doi.org/10.1016/j.scitotenv.2023.162322>
- Tamura K, Stecher G, Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Battistuzzi FU (Ed.). *Molecular Biology and Evolution* 38: 3022–3027. <https://doi.org/10.1093/molbev/msab120>
- Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Bork P, Abarenkov K (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycKeys* 10: 1–43. <https://doi.org/10.3897/mycokeys.10.4852>
- Thomsen PF, Willerslev E (2015) Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation* 183: 4–18. <https://doi.org/10.1016/j.biocon.2014.11.019>

- Tragin M, Vaulot D (2019) Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Scientific Reports* 9: 5190. <https://doi.org/10.1038/s41598-019-41680-6>
- Tremblay J, Singh K, Fern A, Kirton ES, He S, Woyke T, Lee J, Chen F, Dangl JL, Tringe SG (2015) Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology* 6: 771. <https://doi.org/10.3389/fmicb.2015.00771>
- Tringe SG, Rubin EM (2005) Metagenomics: DNA sequencing of environmental samples. *Nature Reviews. Genetics* 6: 805–814. <https://doi.org/10.1038/nrg1709>
- Truelove NK, Patin NV, Min M, Pitz KJ, Preston CM, Yamahara KM, Zhang Y, Raanan BY, Kieft B, Hobson B, Thompson LR, Goodwin KD, Chavez FP (2022) Expanding the temporal and spatial scales of environmental DNA research with autonomous sampling. *Environmental DNA* 4: 972–984. <https://doi.org/10.1002/edn3.299>
- Tsuji S, Miya M, Ushio M, Sato H, Minamoto T, Yamanaka H (2020) Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA* 2: 42–52. <https://doi.org/10.1002/edn3.44>
- Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton ($\leq 3 \mu\text{m}$) in marine ecosystems. *FEMS Microbiology Reviews* 32: 795–820. <https://doi.org/10.1111/j.1574-6976.2008.00121.x>
- Vincent F, Ibarbalz FM, Bowler C (2022) Global marine phytoplankton revealed by the Tara Oceans expedition. In: Clementson LA, Eriksen RS, Willis A (Eds) *Advances in Phytoplankton Ecology*, Elsevier, 531–561. <https://doi.org/10.1016/B978-0-12-822861-6.00024-8>
- Wintzingerode FV, Göbel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: Pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews* 21: 213–229. <https://doi.org/10.1111/j.1574-6976.1997.tb00351.x>
- Worden AZ, Janouskovec J, McRose D, Engman A, Welsh RM, Malfatti S, Tringe SG, Keeling PJ (2012) Global distribution of a wild alga revealed by targeted metagenomics. *Current Biology* 22: R675–R677. <https://doi.org/10.1016/j.cub.2012.07.054>
- Xu Z, Cheung S, Endo H, Xia X, Wu W, Chen B, Ho NHE, Suzuki K, Li M, Liu H (2022) Disentangling the ecological processes shaping the latitudinal pattern of phytoplankton communities in the Pacific Ocean. *mSystems* 7: e01203-21. <https://doi.org/10.1128/msystems.01203-21>
- Yamahara KM, Preston CM, Birch J, Walz K, Marin R, Jensen S, Pargett D, Roman B, Ussler W, Zhang Y, Ryan J, Hobson B, Kieft B, Raanan B, Goodwin KD, Chavez FP, Scholin C (2019) *In situ* autonomous acquisition and preservation of marine environmental DNA using an autonomous underwater vehicle. *Frontiers in Marine Science* 6: 373. <https://doi.org/10.3389/fmars.2019.00373>
- Zhang S, Zhao J, Yao M (2020) A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution* 11: 1609–1625. <https://doi.org/10.1111/2041-210X.13485>
- Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology* 52: 79–92. <https://doi.org/10.1016/j.femsec.2004.10.006>

Supplementary material 1

Sample collection information and associated metadata

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: xlsx

Explanation note: “NA” stands for “not applicable” and “NC” stands for “not collected.” Depth values for samples collected with the 3G-LRAUV-ESP are provided as “INT” for “integrated” since the LRAUV changed depths during the course of filtration. Environmental metadata for the 3G-LRAUV-ESP samples are also generally “NA” for the same reason, and/or because they were not collected by the instrument, except for temperature and salinity which were recorded and averaged over the course of the sampling interval. Units for each category are provided in the column headers; for example, “CHLA_ug_L” indicates chlorophyll a values are provided in µg/L.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl1>

Supplementary material 2

Metagenome and amplicon sequence data set information including per-sample read counts and average read lengths

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: xlsx

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl2>

Supplementary material 3

Pearson's correlation coefficients for phylum-level regressions of relative abundances in metagenome assemblies and corresponding marker gene samples

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: xlsx

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl3>

Supplementary material 4

Reference sequences downloaded from GenBank to build the Magnoliopsida ASV phylogeny

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: xlsx

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl4>

Supplementary material 5

Reference database content for whole genomes and marker genes for each eukaryotic phylum

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: xlsx

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl5>

Supplementary material 6

Rarefaction curves for metagenomes, 18S amplicons, and COI amplicons

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S1**: Rarefaction curves for metagenomes (**A–C**), 18S amplicons (**D–F**), and COI amplicons (**G–I**). Metagenome curves were calculated as a function of k-mers sampled vs unique k-mers detected, while amplicons were calculated as sequences sampled vs unique species detected.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl6>

Supplementary material 7

Krona plots showing the taxonomic breakdown of the Phylum Chordata

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S2**: Krona plots showing the taxonomic breakdown of the Phylum Chordata in (**A**) 18S, (**B**) COI, (**C**) unassembled metagenome reads, and (**D**) metagenome contigs. The percentages of the eukaryotic fraction and the total sequence data for the phylum are provided in the top right of each plot.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl7>

Supplementary material 8

Phylum Dinophyceae correlations

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S3:** Phylum Dinophyceae correlations between paired COI and metagenome sample data and a maximum likelihood phylogeny of all COI ASVs annotated as belonging to the phylum Dinophyceae across all samples. The outlier sample is circled in red on the correlation and plot and the ASV observed at high levels in a single sample is marked by a red box as a sister clade to five other Dinophyceae ASVs.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl8>

Supplementary material 9

Correlations between marker gene and metagenome relative abundances for the Phyla Bacillariophyta, Chlorophyta, and Haptophyta

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S4:** Correlations between marker gene and metagenome relative abundances for the Phyla Bacillariophyta (**A, B**), Chlorophyta (**C, D**), and Haptophyta (**E, F**). Each point represents class-level frequencies in one eDNA sample, with each color representing a different class.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl9>

Supplementary material 10

The levels of Streptophyta annotations in shotgun metagenomes for samples in three depth groupings

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S5:** The levels of Streptophyta annotations in shotgun metagenomes for samples in three depth groupings: surface (0–7 m), deep (10–100 m), and integrated.

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl10>

Supplementary material 11

A maximum likelihood phylogeny of all 18S ASVs annotated as belonging to the family Magnoliopsida (Phylum Streptophyta)

Authors: Nastassia V. Patin, Kathleen J. Pitz, Jacoby D. Baker, Francisco P. Chavez, Kelly D. Goodwin

Data type: png

Explanation note: **figure S6:** A maximum likelihood phylogeny of all 18S ASVs annotated as belonging to the family Magnoliopsida (Phylum Streptophyta). All ASVs showed a more distant relationship to the three reference Magnoliopsida genes than either *Micromonas pusilla* (Phylum Chlorophyta) or *Acrodontium crateriforme* (Phylum Ascomycota).

Copyright notice: This dataset is made available under the Open Database License (<http://opendatacommons.org/licenses/odbl/1.0/>). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: <https://doi.org/10.3897/mbmg.9.144935.suppl11>